AD_____

Award Number: W81XWH-04-1-0058

TITLE: Decipher the Transcriptional Program in Prostate Cancer Cells

PRINCIPAL INVESTIGATOR: Xiaowei Yan, Ph.D.

CONTRACTING ORGANIZATION: Institute for Systems Biology
Seattle, Washington 98103-8904

REPORT DATE: October 2005

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-10-2005	Annual Summary	1 Apr 2004 – 31 Aug 2005

4. TITLE AND SUBTITLE

Decipher the Transcriptional Program in Prostate Cancer Cells

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0058

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Xiaowei Yan, Ph.D.

E-Mail: xyan@systemsbiology.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Systems Biology
Seattle, Washington 98103-8904

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A systematic computational approach is proposed to characterize the transcriptional factor binding sites and related molecular pathways involved in prostate cancer, based on the available expression data generated from systems approaches on prostate cancer cells as well as comparative genomics data from the human and mouse genome sequencing projects. Androgen Response pathways in LNCaP and CL-1 cells are analyzed and re-constructed from MPSS and ICAT experiment datasets, the TRANSFAC database, the Mogul scanner, and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database.

15. SUBJECT TERMS
No subject terms provided

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	17	19b. TELEPHONE NUMBER *(include area code)*

Contents

Introduction

Prostate cancer is the most common non-dermatologic cancer in the United States [1]. It can potentially arise from altered genes in multiple information pathways and followed by a sequential process of mutation along with the progression of the cancer. Therefore, prostate cancer can be taken as a variety of distinct diseases, each of which results from some kind of perturbation of a different information pathway. For example, LNCaP cells are androgen responsive that harbor a functional androgen receptor, while CL-1 cells are androgen-independent that survive an extended period of androgen deprivation [2].

Androgens are steroid hormones, synthesized in the Leydig cells in testis. By binding and activating the androgen receptor (AR) protein, androgens exert their effects in vivo in the development, differentiation, and function of male reproductive and accessory sex tissues such as prostate. Once androgens enter the cell and bind to AR, it undergoes a conformation change in its ligand-binding domain, and causes the dissociation of several accessory proteins. Then the DNA binding domain of the AR can bind to a specific sequence called the androgen responsive element (ARE) which appears in promoter area of many genes with important function in maintaining the male phenotype. This binding of AR dimmer to ARE induces transcriptional activities of those androgen responsive genes. Some of these ARE-containing genes are also

transcriptional factors, like NKX3.1 [3].

The goal of this project is to understand the transcriptional program behind the prostate cancer and thus provide a logical framework of finding novel target genes for developing therapies for prostate cancer. To date, about 20 genes have been confirmed to be transcriptionally regulated by AR through AREs. Although presence of ARE-like sequence near or in a gene does not necessarily imply that it is regulated by AR in vivo, such sequences, particularly when found in regulatory region of a gene, can guide an experimental test of its functional relationship to AR. The consensus ARE is a palindrome, two 6bp (5'-TGTTCT-3') inverted sites separated by a 3bp spacer. Recently, some AR specific-regulated genes are found to have a head-to-tail combination of the hexamer sites, and the right half of the direct repeat is more conserved than the left half. On the other hands, the molecular pathways accounting for androgen regulation remain incompletely characterized. It seems rational to set up a model of the androgen response program involving combinations and interactions between multiple known and unknown pathways. We hypothesize that genes and proteins associated with key nodes within and between special pathways in androgen sensitive and independent cells have substantial potential as therapeutic targets for prostate cancer.

Key Research Accomplishments

1) Comprehensive datasets from our experiments are integrated into robust relational databases, as hosted and shown in SBEAMS (http://db.systemsbiology.net/sbeams/) and LYNX Signome Browser (https://sgb.lynxgen.com/sgb/sgb/).

2) Position weighted matrices (PWM) are generated for the androgen responsive element (ARE) based on known AREs in human genome, which is important for identifying genes with AREs and/or other known TFBSs involved in androgen response program;

3) A set of putative AREs are screened by using Mogul on 5k upstream region of each gene which is significantly up-regulated according to our MPSS (massively parallel signature sequencing) and ICAT (isotope-coded affinity tags) experimental data;

4) Androgen receptor (AR) pathways are reconstructed based on the data from our experiments and other public data resources, such as the TRANSFAC database and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database. These pathways are visualized by using both Cytoscape and BioTapestry tools, providing an approach to identify possible genes and proteins which are associated with key nodes within and between special pathways in

androgen sensitive and independent cells and thus are substantial potential as therapeutic targets for prostate cancer;

5) Comprehensive datasets from our experiments are integrated into a robust relational database, which is also hosted in SBEAMS (http://db.systemsbiology.net/ sbeams/) and LYNX Signome Browser (https://sgb.lynxgen.com/sgb/sgb/) for shared access;

6) 49 tissue-specific MPSS datasets are integrated to analyze tissue-specificity, and various tissue-specific genes are identified in terms of Unigene Cluster ID, including those for prostate, ovary, liver and mammary gland. Tissue-specificity is also analyzed by homolog comparison of human with mouse, based on public mouse MPSS data sources;

7) Secretory prediction, digested peptides, and N-glycosylation sites are calculated for all human Refseq proteins, which is useful for further data analysis and experiment;

8) Some potential biomarkers are identified for further evaluation and validation. This might potentially improve the prediction and diagnoses of prostate cancer;

9) A windows version of stand-alone siRNA designer is programmed and tested. This program can be used to scan the length of the target gene for candidate siRNAs that satisfy the user-specified rules. For each siRNA candidate, a set of deltaG values and score is calculated. Top siRNA candidates are then selected for Blast. Blast files are parsed and blast summary is reported in a summary file along with siRNA candidate hybrid Melting Temperature. NCBI-Blast, WU-Blast and/or Fasta searches can be chosen as blast methods for siRNA candidates.

Reportable Outcomes

Model for disease-perturbed network is set up. Algorithms are designed and implemented to calculate perturbation networks between two stages of prostate cancer progression. Gene information from 314 BioCarta pathways and 155 KEGG pathways is extracted. 37 BioCarta and 14 KEGG pathways are identified up-regulated, and 23 BioCarta and 22 KEGG pathways are found down-regulated, in LNCaP cells versus CL1 cells, based on MPSS data. This is a significant step to understanding prostate cancer progression by means of systems approach.

Conclusions

The systems biology provides a new powerful approach to identifying AR pathways. These pathways are interesting for studies on prostate cancer because the disease process should be reflected in disease-perturbed protein and gene regulatory networks [4].

Abbreviations

AR	Androgen Receptor
ARE	Androgen Receptor Element
ICAT	isotope-Coded Affinity Tags
MPSS	Massively Parallel Signature Sequencing
PWM	Position Weighted Matrix
SBEAMS	Systems Biology Experiment Analysis System

References

1. Greenlee RT, Murray T, Bolden S, Wingo PA. Cancer statistics, 2000. CA Cancer J Clin 2000; 50: 7-33.
2. Patel BJ, Pantuck AJ, Zisman A, et al. CL1-GFP: an androgen independent metastatic tumor model for prostate cancer. J Urol 2000; 164: 1420-5.
3. He WW, Sciavolino PJ, Wing J, et al. A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. Genomics 1997; 43: 69-77.
4. Lin B, White JT, Lu W, et al. Evidence for the presence of disease-perturbed networks in prostate cancer cells by genomic and proteomic analyses: a systems approach to disease. Cancer Res 2005; 65(8): 3081-91.
5. Hood L, Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 2004; 22: 1251-7.
6. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science 2004; 306: 640-3.

Appendices

Publication

1. Lin B, White JT, Lu W, et al. Evidence for the presence of disease-perturbed networks in prostate cancer cells by genomic and proteomic analyses: a systems approach to disease. Cancer Res 2005; 65(8): 3081-91.

Evidence for the Presence of Disease-Perturbed Networks in Prostate Cancer Cells by Genomic and Proteomic Analyses: A Systems Approach to Disease

Biaoyang Lin,[1] James T. White,[1] Wei Lu,[1] Tao Xie,[1] Angelita G. Utleg,[1] Xiaowei Yan,[1] Eugene C. Yi,[1] Paul Shannon,[1] Irina Khrebtukova,[3] Paul H. Lange,[2] David R. Goodlett,[1] Daixing Zhou,[3] Thomas J. Vasicek,[3] and Leroy Hood[1]

[1]Institute for Systems Biology; [2]Department of Urology, University of Washington, Seattle, Washington; and [3]Lynx Therapeutics, Inc., Hayward, California

Abstract

Prostate cancer is initially responsive to androgen ablation therapy and progresses to androgen-unresponsive states that are refractory to treatment. The mechanism of this transition is unknown. A systems approach to disease begins with the quantitative delineation of the informational elements (mRNAs and proteins) in various disease states. We employed two recently developed high-throughput technologies, massively parallel signature sequencing (MPSS) and isotope-coded affinity tag, to gain a comprehensive picture of the changes in mRNA levels and more restricted analysis of protein levels, respectively, during the transition from androgen-dependent LNCaP (model for early-stage prostate cancer) to androgen-independent CL1 cells (model for late-stage prostate cancer). We sequenced >5 million MPSS signatures, obtained >142,000 tandem mass spectra, and built comprehensive MPSS and proteomic databases. The integrated mRNA and protein expression data revealed underlying functional differences between androgen-dependent and androgen-independent prostate cancer cells. The high sensitivity of MPSS enabled us to identify virtually all of the expressed transcripts and to quantify the changes in gene expression between these two cell states, including functionally important low-abundance mRNAs, such as those encoding transcription factors and signal transduction molecules. These data enable us to map the differences onto extant physiologic networks, creating perturbation networks that reflect prostate cancer progression. We found 37 BioCarta and 14 Kyoto Encyclopedia of Genes and Genomes pathways that are up-regulated and 23 BioCarta and 22 Kyoto Encyclopedia of Genes and Genomes pathways that are down-regulated in LNCaP cells versus CL1 cells. Our efforts represent a significant step toward a systems approach to understanding prostate cancer progression. (Cancer Res 2005; 65(8): 3081-91)

Introduction

Prostate cancer is the most common nondermatologic cancer in the United States (1). Initially, its growth is androgen dependent; early-stage therapies, including chemical and surgical castration, kill cancerous cells by androgen deprivation. Although such therapies produce tumor regression, they eventually fail because most prostate carcinomas become androgen independent (2). To improve the efficacy of prostate cancer therapy, it is necessary to understand the molecular mechanisms underlying the transition from androgen dependence to androgen independence.

The transition from androgen-dependent to androgen-independent status likely results from multiple processes, including activation of oncogenes, inactivation of tumor suppressor genes, and changes in key components of signal transduction pathways and gene regulatory networks. Systems approaches to biology and disease are predicated on the identification of the elements of the systems, the delineation of their interactions, and their changes in distinct disease states. Biological information is of two types: the digital information of the genome (e.g., genes and *cis*-control elements) and environmental cues. Normal protein and gene regulatory networks may be perturbed by disease, through genetic and/or environmental perturbations, and understanding these differences lies at the heart of systems approaches to disease. Disease-perturbed networks initiate altered responses that bring about pathologic phenotypes, such as the invasiveness of cancer cells.

To map network perturbations in cancer initiation and progression, one must measure changes in expression levels of virtually all transcripts. Certain low-abundance transcripts, such as those encoding transcription factors and signal transducers, wield significant regulatory influences in spite of the fact they may be present in the cell at very low copy numbers. Differential display (3) or cDNA microarrays (4, 5) have been used to profile changes in gene expression during the androgen-dependent to androgen-independent transition; however, those technologies can identify only a limited number of more abundant mRNAs, and they miss many low-abundance mRNAs due to their low detection sensitivities. Massively parallel signature sequencing (MPSS), a recently introduced method, allows 20-nucleotide signature sequences to be determined in parallel for >1,000,000 DNA sequences from an individual cDNA library or cell state (6). The frequency of each MPSS signature was calculated for each sample and represented in transcripts per million (tpm). MPSS technology allows identification and cataloging of almost all mRNAs, even those with one or a few transcripts per cell. Differentially expressed genes thus identified can be mapped onto cellular networks to provide a systemic understanding of changes in cellular state.

Although transcriptome (mRNA levels) differences are easier to study than proteome (protein levels) differences, cellular functions are usually performed by proteins. RNA expression profiling studies do not address how the encoded proteins function biologically, and

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Biaoyang Lin, Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103. Phone: 206-732-1297; Fax: 206-732-1299; E-mail: blin@systemsbiology.org.

transcript abundance levels do not always correlate with protein abundance levels (7). We therefore complemented our mRNA expression profiling with a more limited protein profiling by using isotope-coded affinity tags (ICAT) coupled with tandem mass spectrometry (MS/MS; ref. 8).

The LNCaP cell line is a widely used androgen-sensitive model for early-stage prostate cancer from which androgen-independent sublines have been generated (4, 5, 9). The cells of one such variant, CL1, in contrast to their LNCaP progenitors, are highly tumorigenic and exhibit invasive and metastatic characteristics in intact and castrated mice (9, 10). Thus, CL1 cells model late-stage prostate cancer. MPSS and ICAT data extracted from these model cell lines can be validated by real-time reverse transcription-PCR (RT-PCR) or Western blot analysis in more relevant biological models (tumor xenografts) and in tumor biopsies.

We conducted a MPSS analysis of ~5 million signatures for the androgen-dependent LNCaP cell line and its androgen-independent derivative CL1. Our database offers the first comprehensive view of the digital transcriptomes of two states of prostate cancer cells and allows us to explore the cellular pathways perturbed during the transition from androgen-dependent to androgen-independent growth. We additionally compared protein expression profiles between LNCaP and CL1 cells using ICAT-MS/MS technology. These are the first steps toward a systems approach to disease through an integrative, systemic understanding of prostate cancer progression at the mRNA, protein, and network levels.

Materials and Methods

Massively parallel signature sequencing analysis. LNCaP and CL1 cells were grown as described by Tso et al. (10). MPSS cDNA libraries were constructed, and individual cDNA sequences were amplified, attached to individual beads, and sequenced as described elsewhere (6). The resulting signatures, generally 20 bases long, were annotated using the then most recently annotated human genome sequence (Human Genome Release hg16, released in November 2003) and the human Unigene (Unigene Build 171, released in July 2004) according to a previously published method (11). We considered only 100% matches between a MPSS signature and a genome signature. We also excluded those signatures that expressed at <3 tpm in both LNCaP and CL1 libraries, as they might not be reliably detected (12). Additionally, we classified cDNA signatures by their positions relative to polyadenylation signals and polyadenylic acid [poly(A)] tails and by their orientation relative to the 5′-3′ orientation of source mRNA. The *Z*-test (13, 14) was used to calculate *P*s for comparison of gene expression levels between the cell lines.

Isotope-coded affinity tag analysis. ICAT reagents were purchased from Applied Biosystems, Inc. (Foster City, CA) Fractionation of cells into cytosolic, microsomal, and nuclear fractions (15), as well as ICAT labeling, MS/MS, and data analyses, were done as described by Han et al. (15). In addition, probability score analysis (16) and Automated Statistical Analysis on Protein Ratio (17) were used to assess the quality of MS spectra and to calculate protein ratios from multiple peptide ratios. Descriptions of these software tools are available at http://regis.systemsbiology.net/software. To compare protein and mRNA expression levels, the Unigene numbers of the differentially expressed proteins were used to find MPSS signatures and their expression levels in tpm. If one Unigene had more than one MPSS signature likely due to alternative terminations, the average tpm of all signatures was taken.

Real-time reverse transcription-PCR. All primers were designed with the PRIMER3 program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) and BLAST searched against the human cDNA and expressed sequence tag (EST) database for uniqueness. Primer sequences and PCR conditions are available on request. Real-time PCR was done on an ABI 7700 machine (Applied Biosystems), and SYBR Green dye (Molecular Probes, Inc., Eugene, OR) was used as a reporter. PCR conditions were designed to give bands of the expected size with minimal primer dimer bands.

Identification of perturbed networks. Genes in the 314 BioCarta and 155 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways or networks (http://cgap.nci.nih.gov/Pathways/) were downloaded and compared with the MPSS data using Unigene IDs as identifiers. If a Unigene ID or an Enzyme Classification number corresponded to multiple signatures potentially due to multiple alternatively terminated isoforms, the tpm counts of the isoforms were combined and then subjected to the *Z*-test (13, 14). Genes with *P*s of ≤0.001 were considered to be significantly differentially expressed. The following criteria were used to identify perturbed networks: a perturbed network must have more than three genes represented on our differentially expressed gene list ($P < 0.001$) and at least 50% of those genes must be up-regulated (an up-regulated pathway) or down-regulated (a down-regulated pathway).

Prediction of secreted proteins. Proteins with signal peptides (classic secretory proteins) were predicted using the same criteria described by Chen et al. (18) with the SignalP 3.0 server (http://www.cbs.dtu.dk/services/SignalP-3.0/) and the TMHMM2.0 server. Putatively nonclassic secretory secreted proteins (without signal peptides) were predicted based on the SecretomeP 1.0 server (http://www.cbs.dtu.dk/services/SecretomeP-1.0/) and required an odds ratio score of >3.0.

Results

Massively parallel signature sequencing analyses of the androgen-dependent LNCaP cell line and its androgen-independent variant CL1. Using MPSS technology, we sequenced 2.22 million signature sequences for LNCaP cells and 2.96 million for CL1 cells. We identified a total of 19,595 unique transcript signatures expressed at levels >3 tpm in at least one of the samples. The signatures were classified into three major categories: 1,093 signatures matched repeat sequences, 15,541 signatures matched unique cDNAs or ESTs, and 2,961 signatures had no matches to any cDNA or EST sequences (but did match genomic sequences). The last category included sequences falling into one of three different categories: signatures representing new transcripts yet to be defined, signatures representing polymorphisms in cDNA sequences (a match of a MPSS sequence to cDNA or EST sequences requires 100% sequence identity), or errors in the MPSS reads. Transcript tags with matches to a cDNA or EST sequence were further classified based on the signatures' relative orientation to transcription direction and their position relative to a polyadenylation site and/or poly(A) tail. We also built a searchable MySQL database (http://www.mysql.com) containing the expression levels (tpm), the genomic locations of the MPSS sequences, the cDNAs or EST matches, and the classification of each signature. A detailed description of the schema for classification is available in Supplementary Table S1. A snapshot of a representative data query is shown in Supplementary Fig. S1.

We first restricted our analysis to those MPSS signatures corresponding to cDNAs with poly(A) tails and/or polyadenylation sites, so that corresponding genes could be conclusively identified. We used the *Z*-test (13, 14) to compare differential gene expression between LNCaP and CL1 cells. Using very stringent *P*s (<0.001), we identified 2,088 MPSS signatures (corresponding to 1,987 unique genes, as some genes have two or more MPSS signatures due to alternative uses of polyadenylation sites) with significant differential expression. Of these, 1,011 signatures (965 genes) were overexpressed in CL1 cells and 1,077 signatures (1,022 genes) were overexpressed in LNCaP cells (Supplementary Table S2). The *Z*-score is related to mRNA abundance in the library. For example,

using a cutoff P of <0.001 in our data set, the expression level in tpm changed from 0 to 26 tpm for the most lowly expressed transcript (>26-fold) but changed from 7,591 and 11,206 tpm for the most highly expressed transcript (1.48-fold).

We randomly selected nine genes from the 1,987 differentially expressed genes identified by our MPSS analysis and compared their changes in expression levels with those obtained by quantitative real-time RT-PCR techniques. We showed that the expression levels of these nine genes changed in the same direction (Table 1). The MPSS expression profiling data were also consistent with the available published data. For example, using RT-PCR, Patel et al. (9) showed that CL1 tumors express barely detectable prostate-specific antigen (PSA) and androgen receptor mRNAs compared with LNCaP cells. Our MPSS results indicated that LNCaP cells expressed 584 tpm of androgen receptor and 841 tpm of PSA; CL1 cells did not express either androgen receptor or PSA (0 tpm in both cases). Freedland et al. found that CD10 expression was lost in CL1 cells compared with LNCaP cells (19); likewise, we found that CD10 was expressed at 0 tpm in CL1 cells but at 56 tpm in LNCaP cells. Using cDNA microarrays, Vaarala et al. (4) compared LNCaP cells and another androgen-independent variant, non-PSA-producing LNCaP line, which is similar to CL1, and identified a total of 56 differentially expressed genes. We found that the expression levels of these 56 genes changed in the same direction (concordant) between LNCaP and CL1 cells and between LNCaP and non-PSA-producing LNCaP cells (data not shown). This identification of 1,987 versus 56 differentially expressed genes, respectively, underscores the striking differences in sensitivity between MPSS and cDNA microarray techniques.

To compare the sensitivity of the MPSS and cDNA microarray procedures, we hybridized cDNA microarrays containing 40,000 human cDNAs to the same LNCaP and CL1 RNAs that we used for MPSS. Three replicate array hybridizations were done. MPSS signatures and array clone IDs were mapped to Unigene IDs for data extraction and comparisons. We found that only those genes expressed at >40 tpm by MPSS could be reliably detected as changing levels by cDNA microarray hybridizations [judged by an expression level twice the SD of the background, a standard cutoff value for microarray data analysis (data not shown)]. This observation is consistent with the 33 to 60 tpm sensitivity of microarrays estimated from the experiment of Hill et al. (20), in which known concentrations of synthetic transcripts were added. In LNCaP and CL1 cells, ~68.75% (13,471 of 19,595) of MPSS signatures (>3 tpm) were expressed at a level below 40 tpm; changes in the levels of these genes will be missed by microarray methods. Many attempts have been made to increase the sensitivity of DNA array technology (21, 22). We have not compared these new improvements against MPSS, but it is clear that there will still be significant differences in the levels of change that can be detected.

Serial analysis of gene expression (SAGE; ref. 23) is another technology for gene expression profiling; like MPSS, it is digital and can generate a large number of signature sequences. However, MPSS (~1 million signatures per sample,) can achieve a much deeper coverage than SAGE (typically ~10,000-100,000 signatures sequenced per sample) at reasonable cost. We compared our MPSS data on LNCaP cells against publicly available SAGE data on LNCaP cells (National Center for Biotechnology Information SAGE database) through common Unigene IDs. The SAGE library GSM724 (total SAGE tags sequenced: 22,721; ref. 24) is derived from LNCaP cells with an inactivated *PTEN* gene; it is the SAGE library most similar to our LNCaP cells. Only 400 (~20%) of our 1,987 significantly differentially expressed genes ($P < 0.001$) had any SAGE tag entry in GSM724. These data illustrate the importance of deep sequence coverage in identifying state changes in transcripts expressed at low-abundance levels.

Functional classifications of genes differentially expressed between LNCaP and CL1 cells. Examination of the Gene Ontology classification of our 1,987 genes revealed that multiple cellular processes have changed during the transition from LNCaP to CL1 cells. The completed list, including Gene Ontology annotations, is shown in Supplementary Table S2. The most interesting groups, categorized by function, are shown in Table 2.

Nineteen differentially expressed proteins are related to apoptosis. Twelve of these are up-regulated in CL1 cells, including the apoptosis inhibitors human T-cell leukemia virus type I binding protein 1 and CASP8 and FADD-like apoptosis regulator. Seven are down-regulated in CL1, including programmed cell death 8 and 5 (apoptosis-inducing factors) and BCL2-like 13 (an apoptosis facilitator). Because CL1 cells have increased expression of apoptosis inhibitors and decreased expression of apoptosis inducers, net inhibition of apoptosis may contribute to their

Table 1. Comparison of MPSS and real-time RT-PCR results

MPSS signature	Name	Genbank accession no.	LNCR/LNCX* ratio by real-time PCR	LNCR/LNCX ratio by MPSS	tpm (LNCR)	tpm (LNCX)
GATCTCAGTTGTAAATA	TSPAN-3	BC000704	0.35	0.28	147	521
GATCTCTTTTCAGAAGT	ITM3	NM_030926	0.28	0.16	22	140
GATCCCTCCAATAAATA	PPP6C	AA574270	0.32	0.00	0	27
GATCACAATAAACGATA	PXMP2	NM_018663	0.32	0.14	5	35
GATCAGATTCACGGACC	PTPRM	NM_002845	11.23	8.87	683	77
GATCAACCTGTGGCTGT	GPT2	BC051364	2.02	4.03	250	62
GATCACAAAATGTTGCC	UGT2B15	NM_001077	0.08	0.05	36	702
GATCACAGAAATGCATA	PKIB	NM_032471	0.43	0.11	35	329
GATCCGGGATGGGAGAC	AQP3	BM968943	1.50	3.56	96	27

*LNCR, LNCaP cells stimulated with androgen; LNCX, LNCaP cells starved of androgen.

Table 2. Examples of differentially expressed genes and their functional classifications

Signatures	LNCaP (tpm)	CL1 (tpm)	Description	Genbank accession no.
Apoptosis related				
GATCAAATGTGTGGCCT	0	3,609	Lectin, galactoside binding, soluble, 1 (galectin 1)	BC001693
GATCATAATGTTAACTA	0	14	Pleiomorphic adenoma gene-like 1 (PLAGL1)	NM_002656
GATCATCCAGAGGAGCT	0	16	Caspase-7, apoptosis-related cysteine protease	U40281
GATCGCGGTATTAAATC	0	15	Tumor necrosis factor receptor superfamily, member 12	U75380
GATCTCCTGTCCATCAG	0	24	IL-1, β	M15330
GATCCCCTTCAAGGACA	1	19	Nudix (nucleoside diphosphate linked moiety X) type motif 1	NM_006024
GATCATTGCCATCACCA	51	278	EST, highly similar to CUL2_human Cullin homologue 2	AL832733
GATCTGAAAATTCTTGG	16	56	CASP8 and FADD-like apoptosis regulator	U97075
GATCCACCTTGGCCTCC	49	149	Tumor necrosis factor receptor superfamily, member 10b	NM_003842
GATCATGAATGACTGAC	118	257	Cytochrome *c*	BC009582
GATCAAGTCCTTTGTGA	299	102	Programmed cell death 8 (apoptosis-inducing factor)	H20713
GATCACCAAAACCTGAT	72	24	BCL2-like 13 (apoptosis facilitator)	BM904887
GATCAATCTGAACTATC	563	146	Apoptosis-related protein APR-3 (APR-3)	NM_016085
GATCCCTCTGTACAGGC	83	13	Unc-13-like (*Caenorhabditis elegans*; UNC13), mRNA	NM_006377
GATCTGGTTGAAAATTG	1,006	49	CED-6 protein (CED-6), mRNA	NM_016315
GATCTCCCATGTTGGCT	86	4	CASP2 and RIPK1 domain containing adaptor with death domain	BC017042
GATCAGAAAATCCCTCT	27	1	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 20, 103 kDa	BC011556
GATCAAGGATGAAAGCT	50	3	Programmed cell death 2	D20426
GATCTGATTATTTACTT	1,227	321	Programmed cell death 5	NM_004708
GATCAAGTCCTTTGTGA	299	102	Programmed cell death 8 (apoptosis-inducing factor)	NM_004208
Cyclins				
GATCCTGTCAAAATAGT	2	47	MCT-1 protein (MCT-1), mRNA	NM_014060
GATCATTATATCATTGG	3	39	Cyclin-dependent kinase inhibitor 2B (CDKN2B)	NM_078487
GATCATCAGTCACCGAA	38	396	Cyclin-dependent kinase inhibitor 2A (p16)	BM054921
GATCGGGGGCGTAGCAT	5	43	Cyclin D1	NM_053056
GATCTACTCTGTATGGG	40	144	Cyclin fold protein 1	BG119256
GATCAGCACTCTACCAC	530	258	Cyclin B1	BM973693
GATCTGGTGTAGTATAT	210	77	Cyclin G2	BM984551
GATCAGTACACAATGAA	642	224	Cyclin G1,	BC000196
GATCTCAGTTCTGCGTT	918	308	CDK2-associated protein 1 (CDK2AP1), mRNA	NM_004642
GATCCTGAGCTCCCTTT	2,490	650	Cyclin I	BC000420
GATCATGCAGTGACATA	15	1	KIAA1028 protein	AL122055
GATCTGTATGTGATTGG	28	1	Cyclin M3	AA489077
Kallikreins				
GATCCACACTGAGAGAG	841	0	KLK3	AA523902
GATCCAGAAATAAAGTC	385	0	KLK4	AA595489
GATCCTCCTATGTTGTT	314	0	KLK2	S39329
CD markers				
GATCAGAGAAGATGATA	0	810	CD213a2, IL-13 receptor, α2	U70981
GATCCCTAGGTCTTGGG	23	161	CD213a1, IL-13 receptor, α1	AW874023
GATCCACATCCTCTACA	0	63	CD33, CD33 antigen (gp67)	BC028152
GATCAATAATAATGAGG	0	151	CD44, CD44 antigen	AL832642
GATCCTTCAGCCTTCAG	0	35	CD73, 5′-nucleotidase, ecto (CD73)	AI831695
GATCTGGAACCTCAGCC	1	50	CD49e, integrin, α5	BC008786
GATCAGAGATGCACCAC	8	122	CD138, syndecan 1	BM974052
GATCAAAGGTTTAAAGT	38	189	CD166, activated leukocyte cell adhesion molecule	AL833702
GATCAGCTGTTTGTCAT	53	295	CD71, transferrin receptor (p90, CD71)	BC001188
GATCGGTGCGTTCTCCT	287	509	CD107a, lysosomal-associated membrane protein 1	AI521424
GATCTACAAAGGCCATG	161	681	CD29, integrin, β1	NM_002211
GATCATTTATTTTAAGC	56	0	CD10 (neutral endopeptidase, enkephalinase)	BQ013520
GATCAGTCTTTATTAAT	150	50	CD107b, lysosomal-associated membrane protein 2	AI459107
GATCTTGGCTGTATTTA	84	1,014	CD59 antigen p18-20	NM_000611
GATCTTGTGCTGTGCTA	408	234	CD9 antigen (p24)	NM_001769
Transcription factors				
GATCAAATAACAAGTCT	0	62	Transcription factor BMAL2	BM854818
GATCTCTATGTTTACTT	0	27	Transcription factor BMAL2	BG163364
GATCCTGACACATAAGA	12	74	Transcription factor BMAL2	BF055294

(Continued on the following page)

Table 2. Examples of differentially expressed genes and their functional classifications (Cont'd)

Signatures	LNCaP (tpm)	CL1 (tpm)	Description	Genbank accession no.
GATCATTTTGTATTAAT	10	61	Transcription factor NRF	BC047878
GATCGTCTCATATTTGC	52	0	Transcriptional coactivator tubedown 100	NM_025085
GATCCCCCTCTTCAATG	0	31	Transcriptional coactivator with PDZ-binding motif	AJ299431
GATCAAATGCTATTGCA	1	55	Transcriptional regulator interacting with the PHS-bromodomain 2	AI126500
GATCTGTGACAGCAGCA	140	35	Transducer of ERBB2, 1	BC031406
GATCAAATCTGTACAGT	239	23	Transducer of ERBB2, 2	AA694240
Annexins and their ligands				
GATCCTGTGCAACAAGA	0	69	Annexin A10	BC007320
GATCTGTGGTGGCAATG	41	630	Annexin A11	AL576782
GATCAGAATCATGGTCT	0	1,079	Annexin A2	BC001388
GATCTCTTTGACTGCTG	210	860	Annexin A5	BC001429
GATCCAAAAACATCCTG	83	241	Annexin A6	AL566871
GATCAGAAGACTTTAAT	0	695	Annexin A1	BC001275
GATCAGGACACTTAGCA	0	2,949	S100 calcium-binding protein A10 (Annexin II ligand)	BC015973
Matrix metalloproteinase				
GATCATCACAGTTTGAG	[redacted]	38	MMP 10 (stromelysin 2)	BC002591
GATCCCAGAGAGCAGCT	0	108	MMP 1 (interstitial collagenase)	BC013118
GATCGGCCATCAAGGGA	0	25	MMP 13 (collagenase 3)	AI370581
GATCTGGACCAGAGACA	0	10	MMP 2 (gelatinase A)	BG332150

greater tumorigenicity. Matrix metalloproteinases (MMP), which degrade extracellular matrix components that physically impede cell migration, are implicated in tumor cell growth, invasion, and metastasis. We found that MMPs 1, 2, 10, and 13 are significantly overexpressed in CL1 cells (Table 2), which may partially explain these cells' aggressive and metastatic behavior.

CD markers are generally localized at the cell surface; some may be associated with prostate cancer (25). We converted all currently identified CD markers (CD1-CD247) from the PROW CD index database (http://www.ncbi.nlm.nih.gov/prow/guide/45277084.htm) to Unigene numbers and used these numbers to identify their signatures and their expression levels. We identified 15 CD markers that are differentially expressed between LNCaP and CL1 cells (Z-score < 0.001; Table 2). Eleven CD markers, including CD213a2 and CD213a1, which encode interleukin (IL)-13 receptors α1 and α2, are up-regulated in CL1 cells; three CD markers, CD9, CD10, and CD107, are down-regulated in these cells (Table 2). Six CD markers went from 0 or 1 to >35 tpm (Table 2), making them good digital or absolute markers or therapeutic targets. These data suggest that carefully selected CD markers may be useful in following the progression of prostate cancer and indeed could serve as potential targets for antibody-mediated therapies (25). Additional functional categories can be seen in Supplementary Table S2.

Delineation of disease-perturbed networks in prostate cancer cells. Genes and proteins rarely act alone but rather generally operate in networks of interactions. Identifying key nodes (proteins) in the disease-perturbed networks may provide insights into effective drug targets. Comparing the genes (proteins) currently available in the 314 BioCarta and 155 KEGG pathway or network (http://cgap.nci.nih.gov/Pathways/) databases with the MPSS data through Unigene IDs, we identified 37 BioCarta and 14 KEGG pathways that are up-regulated and 23 BioCarta and 22 KEGG pathways that are down-regulated in LNCaP cells versus CL1 cells (Table 3). The number of genes whose expression patterns changed in each pathway is listed in Table 3. Each gene along with its expression level in LNCaP and CL1 cells is listed pathway by pathway in our database (ftp://ftp.systemsbiology.net/pub/blin/mpss). Changes in these pathways reveal the underlying phenotypic differences between LNCaP and CL1 cells. For example, multiple networks involved in modulating cell mobility, adhesion, and spreading are up-regulated in CL1 cells, which are more metastatic and invasive than LNCaP cells (Table 3). In the uCalpain and friends in cell spread pathway, calpains are calcium-dependent thiol proteases implicated in cytoskeletal rearrangements and cell migration. During cell migration, calpain cleaves target proteins, such as talin, ezrin, and paxillin, at the leading edge of the membrane while at the same time cleaving the cytoplasmic tails of the integrins β_1(a) and β_3(b) to release adhesion attachments at the trailing membrane edge. Increased activity of calpains increases migration rates and facilitates cell invasiveness (26).

Many pathways we identified as perturbed in the LNCaP and CL1 comparison are interconnected to form networks (in fact, there are probably no discrete pathways, only networks). For example, the insulin signaling pathway, the signal transduction through IL-1 receptor pathway, and nuclear factor-κB (NF-κB) signaling pathway are interconnected through c-Jun, IL-1 receptor, and NF-κB. The mapping of genes onto networks/pathways will be an ongoing objective as more networks/pathways become available. Our transcriptome data will be an invaluable resource in delineating these relationships.

As gene regulatory networks controlled by transcription factors form the top layer of the hierarchy that controls the physiologic network, we sought to identify differentially expressed transcription factors. Of 554 transcription factors expressed in LNCaP and CL1 cells, 112 showed significantly different levels between the cell lines (P < 0.001; Supplementary Table S3). This clearly showed significant difference in the functioning of the corresponding gene regulatory networks during the progression of prostate cancer from the early to late stages.

Table 3. Pathways that are up-regulated or down-regulated comparing LNCaP cells to CL1 cells

Pathways	No. gene hits in a pathway	No. $P < 0.001$ and LNCA > CL1	No. $P < 0.001$ and LNCA < CL1	No. no change
Up-regulated pathways in LNCaP cells				
BioCarta pathways				
Mechanism of gene regulation by peroxisome proliferators via PPARα	35	9	2	24
T-cell receptor signaling pathway	21	6	2	13
ATM signaling pathway	15	5	2	8
CARM1 and regulation of the estrogen receptor	18	5	2	11
HIV type I Nef negative effector of Fas and tumor necrosis factor	33	5	2	26
Epidermal growth factor signaling pathway	17	5	1	11
Role of BRCA1, BRCA2, and ATR in cancer susceptibility	16	5	1	10
Tumor necrosis factor receptor 1 signaling pathway	17	5	1	11
Toll-like receptor pathway	17	5	1	11
FAS signaling pathway CD95	17	4	1	12
Vascular endothelial growth factor hypoxia and angiogenesis	16	4	1	11
Bone remodeling	9	3	1	5
Estrogen receptor–associated degradation ERAD pathway	11	3	1	7
Estrogen-responsive protein Efp controls cell cycle and breast tumors growth	11	3	1	7
Influence of Ras and Rho proteins on G_1-S transition	16	3	1	12
Inhibition of cellular proliferation by Gleevec	13	3	1	9
Mitogen-activated protein kinase inactivation of SMRT corepressor	9	3	1	5
NF-κB activation by nontypeable *Haemophilus influenzae*	16	3	1	12
Rb tumor suppressor checkpoint signaling in response to DNA damage	10	3	1	6
Transcription regulation by methyltransferase of CARM1	10	3	1	6
Ceramide signaling pathway	13	4	0	9
Cystic fibrosis transmembrane conductance regulator and β2-adrenergic receptor pathway	7	4	0	3
Nerve growth factor pathway	11	4	0	7
Platelet-derived growth factor signaling pathway	16	4	0	12
Tumor necrosis factor stress-related signaling	14	4	0	10
Activation of COOH-terminal Src kinase by cyclic AMP-dependent protein kinase inhibits signaling through the T-cell receptor	9	3	0	6
AKAP95 role in mitosis and chromosome dynamics	11	3	0	8
Attenuation of GPCR signaling	7	3	0	4
Chaperones modulate IFN signaling pathway	11	3	0	8
ChREBP regulation by carbohydrates and cyclic AMP	12	3	0	9
Insulin-like growth factor-1 signaling pathway	11	3	0	8
Insulin signaling pathway	11	3	0	8
NF-κB signaling pathway	11	3	0	8
Protein kinase A at the centrosome	12	3	0	9
Regulation of ck1 cdk5 by type 1 glutamate receptors	10	3	0	7
Role of mitochondria in apoptotic signaling	10	3	0	7
Signal transduction through IL-1 receptor	14	3	0	11
KEGG pathways				
Aminosugar metabolism	24	9	4	11
Androgen and estrogen metabolism	37	13	5	19
Benzoate degradation via hydroxylation	5	3	1	1
C21-Steroid hormone metabolism	4	1	0	3
C5-Branched dibasic acid metabolism	2	2	0	0
Carbazole degradation	1	1	0	0
Terpenoid biosynthesis	6	4	1	1
Chondroitin-heparan sulfate biosynthesis	14	8	3	3
Fatty acid biosynthesis (path 1)	3	2	0	1

(Continued on the following page)

Table 3. Pathways that are up-regulated or down-regulated comparing LNCaP cells to CL1 cells (Cont'd)

Pathways	No. gene hits in a pathway	No. $P < 0.001$ and LNCA > CL1	No. $P < 0.001$ and LNCA < CL1	No. no change
Fluorene degradation	3	2	0	1
Pentose and glucuronate interconversions	19	9	1	9
Phenylalanine, tyrosine, and tryptophan biosynthesis	10	5	2	3
Porphyrin and chlorophyll metabolism	28	13	3	12
Streptomycin biosynthesis	6	4	1	1
Up-regulated pathways in CL1 cells				
BioCarta pathways				
Rho cell motility signaling pathway	18	2	6	10
Trefoil factors initiate mucosal healing	14	1	6	7
Integrin signaling pathway	14	1	5	8
Ca^{2+}/calmodulin-dependent protein kinase activation	7	1	4	2
Effects of calcineurin in keratinocyte differentiation	9	1	4	4
Angiotensin II–mediated activation of c-Jun NH_2-terminal kinase pathway via Pyk2-dependent signaling	12	1	3	8
Bioactive peptide-induced signaling pathway	16	1	3	12
CBL-mediated ligand-induced down-regulation of epidermal growth factor receptors	6	1	3	2
Control of skeletal myogenesis by HDAC calcium calmodulin-dependent kinase CaMK	12	1	3	8
How does *Salmonella* hijack a cell	8	1	3	4
Melanocyte development and pigmentation pathway	4	1	3	0
Overview of telomerase protein component gene hTert transcriptional regulation	7	1	3	3
Regulation of PGC-1a	9	0	4	5
ADP-ribosylation factor	9	0	3	6
Down-regulated of MTA-3 in estrogen receptor–negative breast tumors	7	0	3	4
Endocytotic role of NDK phosphins and dynamin	7	0	3	4
Mechanism of protein import into the nucleus	7	0	3	4
Nuclear receptors in lipid metabolism and toxicity	7	0	3	4
Pertussis toxin–insensitive CCR5 signaling in macrophage	9	0	3	6
Platelet amyloid precursor protein pathway	5	0	3	2
Role of Ran in mitotic spindle regulation	8	0	3	5
Sumoylation by RanBP2 regulates transcriptional repression	8	0	3	5
uCalpain and friends in cell spread	5	0	3	2
KEGG pathways				
Arginine and proline metabolism	45	7	16	22
ATP synthesis	31	7	15	9
Biotin metabolism	5	1	3	1
Blood group glycolipid biosynthesis, lactoseries	12	1	6	5
Cyanoamino acid metabolism	5	0	3	2
Ethylbenzene degradation	9	1	3	5
Ganglioside biosynthesis	16	2	6	8
Globoside metabolism	17	3	8	6
Glutathione metabolism	26	4	10	12
Glycine, serine, and threonine metabolism	32	6	14	12
Glycosphingolipid metabolism	35	6	18	11
Glycosylphosphatidylinositol-anchor biosynthesis	26	5	12	9
Glyoxylate and dicarboxylate metabolism	9	1	6	2
Huntington's disease	25	4	10	11
Methane metabolism	9	1	3	5
O-Glycans biosynthesis	19	3	8	8
One-carbon pool by folate	12	2	8	2
Oxidative phosphorylation	93	21	45	27
Parkinson's disease	30	5	14	11
Phospholipid degradation	21	4	12	5
Synthesis and degradation of ketone bodies	7	1	3	3
Urea cycle and metabolism of amino groups	18	2	8	8

As secreted proteins can readily be exploited for blood cancer diagnosis and prognosis, we next asked how many of our differentially expressed genes encode secreted proteins. We identified 521 signatures belonging to 460 genes potentially encoding secreted proteins (Supplementary Table S6). Among these, 287 (259 genes) and 234 (201 genes) signatures, respectively, are overexpressed or underexpressed in CL1 cells compared with LNCaP cells. Thus, one can think about using blood diagnostics (changes in relevant protein concentrations) to follow prostate cancer progression.

Quantitative proteomic analysis of prostate cancer cells. We quantitatively profiled the protein expression changes between LNCaP and CL1 cells using the ICAT-MS/MS protocol described by Han et al. (15). We generated a total of 142,849 MS/MS, 7,282 of which corresponded to peptides with a mass spectrum quality score *P* of >0.9 (allowing unambiguous identification of peptides; ref. 16). We obtained quantitative peptide ratios for 4,583 peptides corresponding to 940 proteins. The number of peptides is greater than the number of proteins because (*a*) mass spectrometry identified multiple peptides from the same protein and (*b*) the ionization step of mass spectrometry created different charge states for the same peptide. The protein ratios were calculated from multiple peptide ratios using an algorithm for the Automated Statistical Analysis on Protein Ratio (17). In the end, we identified 82 proteins that are down-regulated and 108 proteins that are up-regulated by at least 1.8-fold in LNCaP cells compared with CL1 cells. The functional classification of the proteins identified is shown in Supplementary Table S4.

Fifty-four percent (103 of 190) of differentially expressed proteins identified have enzymatic activity. Many of the proteins identified are involved in fatty acid and lipid metabolism, including fatty acid synthase, carnitine palmitoyltransferase II, and propionyl CoA carboxylase α polypeptide. Fatty acid and lipid metabolism is perturbed in prostate cancer (27, 28). Additionally, many genes involved in lipid transport were altered, including five Annexin family proteins, prosaposin, and fatty acid binding protein 5 (Supplementary Table S4). Annexin A1 was shown to be overexpressed in non-PSA-producing LNCaP cells compared with PSA-producing LNCaP cells (4). Annexin A7 is postulated to be a prostate tumor suppressor gene (29). Annexin A2 expression is reduced or lost in prostate cancer cells, and its re-expression inhibits prostate cancer cell migration (30).

Other genes we identified here have been implicated in carcinogenesis, including tumor suppressor p16 and insulin-like growth factor-II receptor (27, 31). Some genes have been implicated previously in prostate cancer, such as prostate cancer overexpressed gene 1 (*POV1*), which is overexpressed in prostate cancer (32), and δ1 and α1 catenin (cadherin-associated protein) and junction plakoglobin, which are down-regulated in prostate cancer cells (33). However, the potential relationships of most of the proteins identified here to prostate cancer require further elucidation. For example, transmembrane protein 4 (*TMEM4*), a gene predicted to encode a 182–amino acid type II transmembrane protein, is down-regulated ~2-fold in CL1 cells compared with LNCaP cells. MPSS data also indicated that TMEM4 is down-regulated ~2-fold in CL1 cells. Many type II transmembrane proteins, such as TMPRSS2, are overexpressed in prostate cancer patients (34). It will be interesting to see whether TMEM4 overexpression plays a primary role in prostate carcinogenesis. We also identified 12 proteins that have not been annotated or functionally characterized. The relationships between these novel proteins and prostate cancer also need further study.

Additionally, we sought to compare the changes in expression at the protein level in the two cell states with changes at the mRNA level. We converted the protein IDs and MPSS signatures to Unigene IDs to compare the MPSS data with the ICAT-MS/MS data. We limited this comparison to those with common Unigene IDs and with reliable ICAT ratios (SD <0.5) and ended up with a subset of 79 proteins. Of these, 66 genes (83.5%) were concordant in their changes in mRNA and protein levels of expression and 13 genes (16.5%) were discordant (i.e., having higher protein expression but lower mRNA expression or vice versa). The scatter plot of protein/mRNA expression ratios is shown in Fig. 1. There are no functional similarities among the discordant genes. As these mRNAs and proteins are expressed at relatively high levels, discordance due to measurement errors is unlikely. Clearly, post-transcriptional mechanism(s) of protein expression is important, although the elucidation of the specific mechanism(s) awaits further studies.

Discussion

The systems approach to disease is predicated on the idea that the disease process is reflected in disease-perturbed protein and gene regulatory networks. Molecular systems biology has two important features: (*a*) it employs global analyses where global implies studying changes in transcript or protein levels as well as the relationships of all of the elements in the system and (*b*) it integrates different types of biological information (single nucleotide polymorphisms, DNA, mRNA, protein, protein interactions, etc.). MPSS is a powerful and sensitive technology that allows deep analysis of the prostate transcriptome. The MPSS protocol we used for this study relies on GATC enzymatic sites to cleave the 3′ region of cDNAs to generate DNA fragments as substrates for MPSS. cDNAs lacking GATC in their 3′ region would be excluded from these analyses. The estimated percentage of cDNA clones lacking an appropriately positioned GATC site is ~3% as calculated from the Mammalian Genome Collection full-length sequences. Among the 15,064 Mammalian Genome Collection sequences, 14,602 (96.93%) sequences have appropriate GATC sites. The protocol we used is also biased toward capturing MPSS signatures within 500 bp 5′ of the poly(A) site. If the GATC site is located beyond 500 bp 5′ of the poly(A) site, it will likely be missed as well. For example, *NKX3.1*, a prostate-specific and androgen-regulated gene (35, 36), is not found in our MPSS data set because its GATC site closest to the poly(A) tail (Genbank accession no. AF247704) is 2.8 kb away. Recently, a new protocol that eliminates this bias was developed at Lynx.[4] We estimated that LNCaP cells expressed ~280,000 transcripts per cell. We obtained ~900 μg of total RNA from 10^8 cells. With an average of 3% polyadenylated RNA and an average transcript length of 1 kb, this corresponds to 280,000 transcripts per cell. Therefore, with >2 million signatures obtained for each cell state by MPSS, we can detect transcripts expressed at levels of <1 transcript per cell (this means that not all cells express the transcript).

[4] D. Zhou, personal communication.

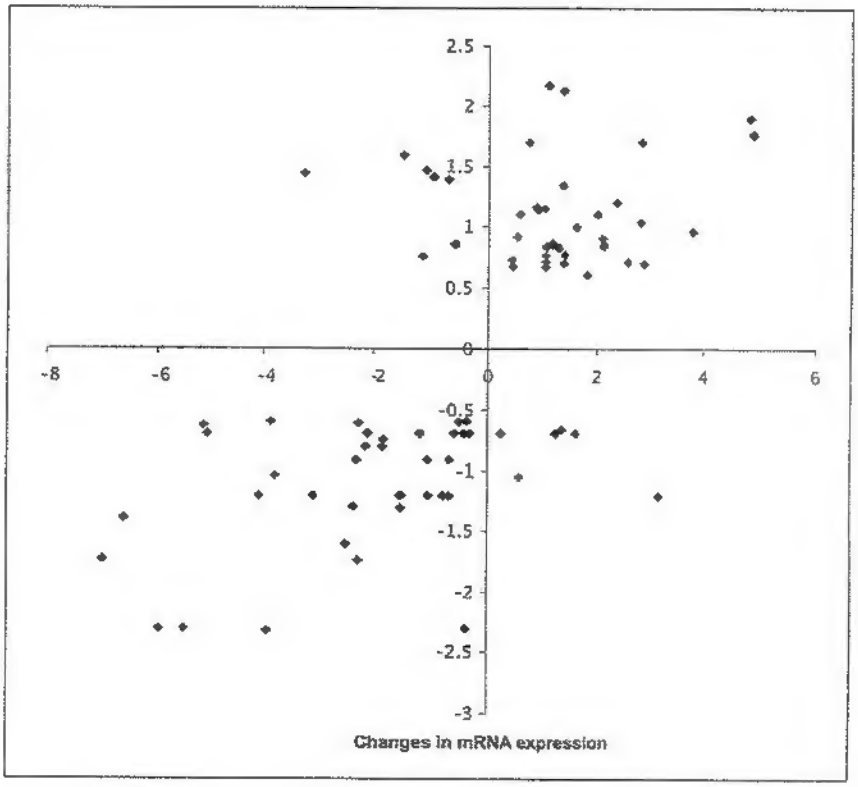

Figure 1. Scatter plot of the protein ratios obtained by ICAT and the mRNA expression ratios obtained by MPSS. Expression ratios in Supplementary Table S3 were transformed to natural logarithms and then plotted.

The BioCarta and KEGG databases describe 469 protein pathways or networks (http://cgap.nci.nih.gov/Pathways/). We have identified 37 BioCarta and 14 KEGG pathways that are up-regulated and 23 BioCarta and 22 KEGG pathways that are down-regulated in LNCaP cells versus CL1 cells. We have also shown that 112 transcription factors change between these two disease states, consistent with the fact that several different gene regulatory networks are perturbed. These changes indicate significant alterations of the corresponding gene regulatory networks. These transcription factors include androgen receptor along with other six transcription factors, such as the ets homologous factor, a liver-specific bHLH-Zip transcription factor, an IFN regulatory factor, and CCCTC-binding factor (zinc finger protein; by exploring data in Supplementary Table S3). The fascinating question is which of these networks are directly correlated with prostate cancer progression and which are changed secondarily as a consequence of their connections to the primary disease networks. We are working on strategies to distinguish these possibilities. Nevertheless, we can firmly conclude that the progression from early-stage to late-stage prostate cancer as represented by LNCaP and CL1 cells clearly is reflected in significant changes in both protein and gene regulatory networks.

In contrast to the MPSS technology, the ICAT technology is an immature technology that cannot now carry out global analyses (37). The integration of different types of data provides powerful new approaches to defining more precisely protein and gene regulatory networks (38). We have shown that the protein and RNA expression levels of 66 of 79 genes (83.5%) were concordant (i.e., changes in the same direction; Supplementary Table S5). This concordance rate is higher than that reported elsewhere (39, 40). Waghray et al. found that only 8 of 25 (32%) androgen-responsive genes in LNCaP cells showed concordance between protein levels measured by two-dimensional gels and MS/MS and mRNA levels analyzed by SAGE (39). Although genes in different experimental systems may have different concordance rates between mRNA and protein expression, use of different methods for quantitative protein profiling (ICAT-MS/MS versus two-dimensional gel-MS/MS) and mRNA expression profiling (MPSS versus SAGE) may also account for the differences. It is also critical to use only those data with high confidence levels in the comparisons between mRNA and protein levels. The expression levels obtained by MPSS are more accurate than those obtained by SAGE or DNA microarrays because of the deep sequence coverage MPSS achieves. We have also limited our data set to only those proteins (649 of them) that were identified in multiple peptide hits and in which the ICAT ratios did not vary greatly among different peptides from the same protein (SD < 0.5). Such variation could derive from experimental errors or from different protein isoforms. There are a multiplicity of post-transcriptional mechanisms that have been described and there are probably more to be identified (41). The important point is that this major aspect of control could not have been identified without the integration of two data types—mRNAs and proteins.

The systems approach provides powerful new approach to diagnostics. The idea is that disease-perturbed networks change their patterns of mRNA and protein expression both within the diseased cells and in terms of the proteins they synthesize that are secreted into the blood. Of the 1,987 mRNAs that changed in the transition from LNCaP to CL1 cells (early-stage to late-stage

cancer), 460 (23.2%) encoded proteins that were potentially secreted (Supplementary Table S6). Sixteen of these putative secreted proteins were also identified to be differentially expressed in these two cell states by the ICAT approach (Supplementary Table S6). Of the 190 differentially expressed proteins identified by the ICAT approaches, 22 were predicted to be secreted proteins (Supplementary Table S6). These proteins are excellent candidates for investigation as diagnostic markers for prostate cancer progression. The interesting point is that these secreted diagnostic markers will serve as surrogates for the state of the corresponding protein and gene regulatory networks and potentially will enable one to (*a*) stratify disease into distinct categories (e.g., relatively benign, slowly invasive, and rapidly metastatic for prostate cancer), for these different types of prostate cancer will employ different disease-perturbed networks); (*b*) follow progression; (*c*) follow response to therapy; and (*d*) monitor adverse drug reactions. The other interesting possibility is that the perturbed secreted proteins will serve as markers to identify the primary disease-perturbed networks and accordingly will identify networks that may harbor excellent protein candidates for drug targeting—drug targets that may kill disease cells specifically or return the networks to a more normal state.

Interestingly, these two states of prostate cancer progression can lead to "digital changes" (i.e., changes from 0 to ≥50 tpm). Thus, one can possibly obtain diagnostic markers that are digital in the sense that they transition from no expression to some expression. In the transition from LNCaP cells to CL1 cells, there are 175 signatures (169 mRNAs) that go from 0 to ≥50 tpm. Likewise, in going from CL1 cells to LNCaP cells, there are 131 signatures (128 mRNAs; Supplementary Table S2). Among the transcription factors we identified, eight transcription factors changed from 0 tpm in LNCaP to >50 tpm in CL1 cells and seven transcription factors changed from >50 tpm in LNCaP cells to 0 tpm in CL1 cells (Supplementary Table S3). Eight pathways were affected by the "digital changes" (Supplementary Table S7). For example, acid ceramidase 1 and aspartate aminotransferase changed from >50 tpm in LNCaP cells to 0 tpm in CL1 cells, affecting multiple pathways, including the insulin-like growth factor-I receptor pathway and activation of COOH-terminal Src kinase pathway (Supplementary Table S7). It will be interesting to test these potential digital diagnostic markers.

Our analyses provide an excellent database and powerful resource enabling the development of tools for multivariable diagnosis and prognosis. They represent a significant step toward a system-wide understanding of prostate cancer progression. The systems approach to disease will offer powerful to approaches to diagnostics, therapeutics, and even prevention in the future (42). It will almost certain usher in an era of predictive and preventive medicine over the next 10 to 20 years (43).

Acknowledgments

Received 11/19/2004; revised 1/24/2005; accepted 2/8/2005.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Greenlee RT, Murray T, Bolden S, Wingo PA. Cancer statistics, 2000. CA Cancer J Clin 2000;50:7–33.
2. Isaacs JT. The biology of hormone refractory prostate cancer. Why does it develop? Urol Clin North Am 1999;26:263–73.
3. Bussemakers MJ, van Bokhoven A, Verhaegh GW, et al. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res 1999;59:5975–9.
4. Vaarala MH, Porvari K, Kyllonen A, Vihko P. Differentially expressed genes in two LNCaP prostate cancer cell lines reflecting changes during prostate cancer progression. Lab Invest 2000;80:1259–68.
5. Chang GT, Blok LJ, Steenbeek M, et al. Differentially expressed genes in androgen-dependent and -independent prostate carcinomas. Cancer Res 1997;57:4075–81.
6. Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000;18:630–4.
7. Chen G, Gharib TG, Huang CC, et al. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 2002;1:304–13.
8. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 1999;17:994–9.
9. Patel BJ, Pantuck AJ, Zisman A, et al. CL1-GFP: an androgen independent metastatic tumor model for prostate cancer. J Urol 2000;164:1420–5.
10. Tso CL, McBride WH, Sun J, et al. Androgen deprivation induces selective outgrowth of aggressive hormone-refractory prostate cancer clones expressing distinct cellular and molecular properties not present in parental androgen-dependent cancer cells. Cancer J Sci Am 2000;6:220–33.
11. Meyers BC, Tej SS, Vu TH, et al. The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. Genome Res 2004;14:1641–53.
12. Jongeneel CV, Iseli C, Stevenson BJ, et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. Proc Natl Acad Sci U S A 2003;100:4702–5.
13. Man MZ, Wang X, Wang Y. POWER_SAGE: comparing statistical tests for SAGE experiments. Bioinformatics 2000;16:953–9.
14. Kal AJ, van Zonneveld AJ, Benes V, et al. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol Biol Cell 1999;10:1859–72.
15. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol 2001;19:946–51.
16. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
17. Li XJ, Zhang H, Ranish JA, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem 2003;75:6648–57.
18. Chen Y, Yu P, Luo J, Jiang Y. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. Mamm Genome 2003;14:859–65.
19. Freedland SJ, Seligson DB, Liu AY, et al. Loss of CD10 (neutral endopeptidase) is a frequent and early event in human prostate cancer. Prostate 2003;55:71–80.
20. Hill AA, Hunter CP, Tsung BT, Tucker-Kellogg G, Brown EL. Genomic analysis of gene expression in *C. elegans*. Science 2000;290:809–12.
21. Han M, Gao X, Su JZ, Nie S. Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. Nat Biotechnol 2001;19:631–5.
22. Bao P, Frutos AG, Greef C, et al. High-sensitivity detection of DNA hybridization on microarrays using resonance light scattering. Anal Chem 2002;74:1792–7.
23. Velculescu VE, Vogelstein B, Kinzler KW. Analysing uncharted transcriptomes with SAGE. Trends Genet 2000;16:423–5.
24. Lal A, Lash AE, Altschul SF, et al. A public database for gene expression in human cancers. Cancer Res 1999;59:5403–7.
25. Liu AY, True LD, LaTray L, et al. Analysis and sorting of prostate cancer cell types by flow cytometry. Prostate 1999;40:192–9.
26. Perrin BJ, Huttenlocher A. Calpain. Int J Biochem Cell Biol 2002;34:722–5.
27. Pandian SS, Eremin OE, McClinton S, Wahle KW, Heys SD. Fatty acids and prostate cancer: current status and future challenges. J R Coll Surg Edinb 1999;44:352–61.
28. Fleshner N, Bagnell PS, Klotz L, Venkateswaran V. Dietary fat and prostate cancer. J Urol 2004;171:S19–24.
29. Cardo-Vila M, Arap W, Pasqualini R. Is Annexin 7 a tumor suppressor gene in prostate cancer? Pharmacogenomics J 2001;1:92–4.
30. Liu JW, Shen JJ, Tanzillo-Swarts A, et al. Annexin II expression is reduced or lost in prostate cancer cells and its re-expression inhibits prostate cancer cell migration. Oncogene 2003;22:1475–85.
31. Chi SG, deVere White RW, Muenzer JT, Gumerlock PH. Frequent alteration of CDKN2 (p16(INK4A)/MTS1) expression in human primary prostate carcinomas. Clin Cancer Res 1997;3:1889–97.
32. Cole KA, Chuaqui RF, Katz K, et al. cDNA sequencing and analysis of POV1 (PB39): a novel gene up-regulated in prostate cancer. Genomics 1998;51:282–7.
33. Kallakury BV, Sheehan CE, Winn-Deen E, et al. Decreased expression of catenins (α and β), p120 CTN, and E-cadherin cell adhesion proteins and E-cadherin gene promoter methylation in prostatic adenocarcinomas. Cancer 2001;92:2786–95.
34. Vaarala MH, Porvari K, Kyllonen A, Lukkarinen O,

Vihko P. The TMPRSS2 gene encoding transmembrane serine protease is overexpressed in a majority of prostate cancer patients: detection of mutated TMPRSS2 form in a case of aggressive disease. Int J Cancer 2001;94:705–10.
35. Prescott JL, Blok L, Tindall DJ. Isolation and androgen regulation of the human homeobox cDNA, NKX3.1. Prostate 1998;35:71–80.
36. Bhatia-Gaur R, Donjacour AA, Sciavolino PJ, et al. Roles for Nkx3.1 in prostate development and cancer. Genes Dev 1999;13:966–77.
37. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.
38. Baliga NS, Bonneau R, Facciotti MT, et al. Genome sequence of *Haloarcula marismortui*: a halophilic archaeom from the Dead Sea. Genome Res 2004;14:2221–34.
39. Waghray A, Feroze F, Schober MS, et al. Identification of androgen-regulated genes in the prostate cancer cell line LNCaP by serial analysis of gene expression and proteomic analysis. Proteomics 2001;1:1327–38.
40. Chen G, Gharib TG, Huang CC, et al. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 2002;1:304–13.
41. Rajagopalan LE, Malter JS. Regulation of eukaryotic messenger RNA turnover. Prog Nucleic Acid Res Mol Biol 1997;56:257–86.
42. Hood L, Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 2004;22:1215–7.
43. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science 2004;306:640–3.